Meta-Psychology, 2024, vol 8, MP.2022.3236
https://doi.org/10.15626/MP.2022.3236
Article type: Original Article

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Rickard Carlsson
Reviewed by: Florian Pargent, Marcus Crede
Analysis reproduced by: Jens Fust
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/2GTHZ

Measurements of Susceptibility to Anchoring are Unreliable: Meta-Analytic Evidence From More Than 50,000 Anchored Estimates

Lukas Röseler[1, 2], Lucia Weber[1], Katharina A. C. Helgerth[1], Elena Stich[1], Miriam Günther[1], Paulina Tegethoff[1], Felix S. Wagner[1], and Astrid Schütz[1]
[1]University of Bamberg
[2]Münster Center for Open Science, University of Münster

Theories on anchoring effects—the assimilation of numerical estimates toward previously considered numbers—have been used to derive hypotheses that susceptibility to anchoring is correlated with certain personality traits. Thus, for the last decade, a considerable amount of research has investigated relationships between people's susceptibility to anchoring and personality traits (e.g., intelligence, the Big Five, narcissism). However, many of the findings are contradictory. We suspect that this inconsistency is grounded in imprecise measurements. Unfortunately, few reports have disclosed estimates of the susceptibility scores' reliability (e.g., Cronbach's Alpha or interitem correlations). We created a large and open data set of anchoring susceptibility scores and conducted a meta-analysis to test how extensive the *reliability problem* is. Results suggest that the reliability of most tasks is very low. In the few cases in which the reliability is acceptable, the validity of the anchoring scores is questionable. We discuss requirements for further attempts to solve the reliability problem.

Keywords: anchoring effect, meta-analysis, reliability, open data, personality, moderator

Is the Rio Grande longer or shorter than 5,000 miles? And is the average number of babies born each day in the United States more or fewer than 100? These are two common examples of questions used to investigate anchoring effects. If such an anchor is provided, later estimates of the answers to both questions are typically biased toward the anchors (i.e., 5,000 miles and 100 babies) and away from the correct values (i.e., 1,885 miles and 10,267 babies). It has been assumed that some people are more susceptible to anchoring effects than others—and this assumption is plausible. For example, a person with little general knowledge, high openness to experience, or low scores on narcissism might stick closer to the anchor than other people. In the present research, we aim to answer the questions of whether, and if so, under which conditions such susceptibility scores resulting from different anchoring items are correlated, that is, when such an overall tendency can be *measured reliably*.

Why Do We Need to Measure Susceptibility to Anchoring Reliably?

Anchoring models suggest that there are person parameters that can explain differences in the susceptibility to anchoring. For example, in the Selective Accessibility Model (Mussweiler & Strack, 1999), susceptibility to priming is equated with susceptibility to anchoring, and in the Insufficient Adjustment Model (e.g., Epley and Gilovich, 2001), adjustment strength is considered a person parameter (e.g., Epley and Gilovich, 2001). Stable individual differences in anchoring are at the core of anchoring theories, and if such trait parameters were to be found, this would advance theory development. Specifically, a key finding that has been cited in support of the insufficient adjustment model is: Need for cognition was correlated with people's susceptibility to anchoring (Epley and Gilovich, 2006, Study 2a), but note that this could not be replicated (Röseler et al., 2022) Similarly, in the field of advice-taking, the behavioral measure of the weight of a person's advice—toward which previous judgments were adjusted after the advice was given—was theoretically and empirically linked to agency (Schultze et al., 2018, p. 12), but the effect size was extremely small (r = -.11, 95% CI [-.22, -.01]).

The reliable and valid measurement of person parameters in anchoring paradigms is thus an essential aspect for providing evidence of the validity of the respective theories: For example, if priming is a valid

account of anchoring, susceptibility to priming (e.g., Smeesters et al., 2009) and susceptibility to anchoring should be correlated, and moderators of priming should also be moderators of anchoring. By contrast, the absence of reliable individual differences in the susceptibility to anchoring would suggest that anchoring is a phenomenon that occurs to similar extents for different people and that other theoretical models, such as scale distortion (Frederick & Mochon, 2012; Mochon & Frederick, 2013), are superior (cf. Bahník, 2021b) to the alternative accounts mentioned earlier.

Moreover, reliable individual differences would allow for an integration of or discrimination between different biases and tasks. Past research has suggested that biases, such as hindsight bias and anchoring (Pohl et al., 2003), are related. However, without reliable measurement, it is not possible to answer questions about whether and how such phenomena are related. Without reliable measurement, it is difficult to determine whether there are different mechanisms that underlie different anchoring effects, such as self-generated anchors and experimenter-provided anchors (e.g., Epley and Gilovich, 2001).

What is the Evidence for a Problem With the Reliability of Anchoring?

The lack of reports on reliability is a symptom of the neglect of reliability issues (Parsons et al., 2018) as has been noted for the dot-probe and Stroop tasks. Similarly, Hedge et al. (2018) discussed the reliability paradox—the observation that large and robust effects often show very low reliability. In anchoring research, there have been few reports on the reliability of anchoring tasks. To our knowledge, initial evidence of low reliability was presented by Röseler et al. (2019). Subsequently, the reliability problem was corroborated for a number of different paradigms (Röseler, 2021; Schindler et al., 2021). Recently, Berthet (2021) was the first to provide evidence that it is possible to measure the susceptibility to confirmation bias reliably.

There are a host of findings on significant moderator variables (e.g., narcissism and autistic tendencies, Cheek and Norem, 2022; cognitive ability, Bergman et al., 2010; Teovanović, 2019). However, meta-analyses (Röseler, 2021) and replication attempts (Cheek & Norem, 2019) suggest that the average effects of many moderators (i.e., Big Five, cognitive ability, cognitive reflection, self-control) are zero.

A small line of research has reported that the susceptibility to anchoring effects can be measured reliably. Among the first cases of reports of the reliability of anchoring tasks is Teovanović (2019; see also Berthet, 2021; Gertner et al., 2016). The common thread across these studies is that they do not use a classical anchoring paradigm but instead rely on the so-called judge-advisor system (e.g., Bonaccio and Dalal, 2006) in which participants (a) give an estimate (i.e., "unanchored estimate"), (b) are given advice (i.e., an anchor), and (c) adjust their estimate. By contrast, in classical anchoring tasks, participants do not provide an estimate before they are given the anchor. Thus, it is possible that this unanchored estimate introduces a second anchor. If participants then provide an estimate that is closer to either of the two values (i.e., unanchored estimate or advice), this tendency could be interpreted as susceptibility to anchoring. In this paradigm, how far participants adjust from their initial value (their "own anchor") toward the advice (the "provided anchor") is reliable (see https://osf.io/t7ckr for analyses of reliabilities for multiple advice-taking studies conducted by Schultze et al., 2017).

Solving the Reliability Problem

If some anchoring tasks do not provide reliable scores, this does not necessarily mean that there are no tasks that provide such scores. For example, reports of reliability might simply have been omitted in cases in which they were acceptable (cf. Parsons et al., 2018). Suggesting that there is at least one anchoring task where the susceptibility to anchoring can be measured reliably is an existential quantification and is similar to "black swans exist." Disproving such a statement is practically impossible, as all instances of swans would have to be observed. By contrast, the truth of such a statement can be established by finding a single black swan (given that method biases or design flaws have been ruled out as alternative explanations). To allow for a reasonable assessment of whether reliable measurement is possible, we analyzed data from existing anchoring studies that each included multiple anchoring items.

Method

Testing the reliability of all possible paradigms of anchoring is not feasible as there are almost no two cases in which the exact same anchoring task was used. For example, paradigms differ with respect to the extremeness Chapman and Johnson (1994) or precision of anchors (e.g., Janiszewski and Uy, 2008), whether participants were asked a comparative question (e.g., "Is the correct value more or less than the anchor?") or not (e.g., "Hint: the correct value is less than the anchor"), whether anchors are random (e.g., a number drawn from a fortune wheel; Tversky and Kahneman, 1974, p. 1128) or plausible estimates (e.g., Mussweiler et al., 2000, p. 1145), and many more

paradigm features. Instead of collecting new data, we collected and aggregated these data sets. The collection is available as part of the Open Anchoring Quest (OpAQ; https://metaanalyses.shinyapps.io/OpAQ), which is a community-augmented meta-analysis on anchoring effects (CAMA; Tsuji et al., 2014). The corresponding PRISMA checklist is available online (https://osf.io/zxenw).

Literature Search

We searched for published articles and preprints on anchoring effects to identify openly accessible data sets that could be used to test for anchoring effects. The inclusion criterion for our *data set* was that a study needed an anchor manipulation with at least one high and one low anchor (within- or between-subjects) and anchored estimates. For example, Cheek and Norem (2022), meeting this inclusion criterion, let participants estimate the year the telephone was invented after having considered either the year 1830 as an anchor or 1915 (the correct year is usually referred to as Phillip Reis' "Telephon" presentation in 1861). For data to be included in our analyses, at least two anchoring trials with true values or unanchored mean estimates (in other words, anchored estimates for at least two anchoring items per participant) were required. Cheek and Norem (2022) also asked participants about other values such as the maximum speed of a house cat in their study, leading to the inclusion of the dataset in the reliability analyses. We chose a more inclusive criterion for the data set (i.e., that datasets with only a single anchoring item are included as well) so that the dataset can be used to study anchoring effect questions beyond the analyses of reliability discussed here (e.g., for a meta-analysis see Röseler and Schütz, 2022). Studies from the judge-advisor-system paradigm were not included because they have already been shown to yield reliable scores, and these scores cannot be interpreted as indicating a susceptibility to anchoring. Sticking to the advice is different from the typical susceptibility to anchoring because the initial "unanchored" judgment and the advice both serve as anchors. The judge-advisor system can be used to measure susceptibility to external advice over internal estimates (i.e., the weight of advice) but not susceptibility to anchors (which can be internally or externally generated).

We searched for publications with openly available data sets via a classic literature search with the keywords anchoring, anchoring effect, anchor effect, scale anchor, anchor precision, and anchor moderator via EBSCOhost, Web of Science, and OSF preprints in several different databases in all available years (see Table 1 for an overview). This yielded 17 research articles. To complement this search, (a) we included five data sets from personal correspondence with other researchers, (b) we emptied our file drawer with studies on anchoring (16 studies, ten of which met the inclusion criterion), (c) we contacted all authors whose data we used and asked them to provide us with additional published or unpublished data (personal call for data), and (d) we issued open calls for data via the Biennial Conference of the German Psychological Society – Personality Psychology and Psychological Diagnostics (DPPD) Section (14 September 2021), Researchgate (21 September, 2021), Facebook's Psychological Methods Discussion Group (6 December, 2021) and PsychMAP group (7 December, 2021), and via mailing lists from the society for personality and social psychology, the society for experimental social psychology, the German Psychological Society (DGPS), and the Society for the Psychological Study of Social Issues (all 6 and 7 December, 2021; see also https://osf.io/b6fny; 12 research articles). We updated this list due to a literature alert with one further dataset. Due to some articles comprising multiple studies, the final number of studies was 96 of which 41 could be included in our analyses as they used more than one anchoring item (see Figure 11 for a PRISMA flow chart).

Coded Data

Eligible data sets were reshaped into a long format where each line represented one anchoring trial. We coded (a) ID variables (e.g., participant identifiers), (b) references, (c) links to the data sets, if available, (d) demographics (i.e., age and gender), and (e) information about the paradigm (e.g., anchoring item, true value or unanchored mean estimate, anchor, estimate, task type, whether the direction of adjustment was known, whether a comparative question was used, where the study was run, how the target stimuli were described, what scale participants used for their estimates). Finally, we coded whether the data were part of a public research article (e.g., peer-reviewed article or preprint) and whether the study had been preregistered. An overview of all the coded variables and their labels is available online (https://osf.io/mdgze).

Susceptibility to anchoring is not equal to the estimate provided by the participant (e.g., the length of the Rio Grande), and many different ways to compute susceptibility scores have been suggested. Susceptibility to anchoring is derived from the estimate, the anchor, and—in some cases—the true value of the respective item. Using theoretical considerations and recommendations from other anchoring researchers, we computed four different scores to operationalize susceptibility to anchoring. First, we used adjustment from the anchor

Table 1

Overview of Searched Databases, Keywords, and Dates of Literature Searches

Search tool	Keywords	Fields	Years	Databases
EBSCOhost	"anchoring effect"	All	1974-2021	ERIC, MEDLINE, PsycInfo, PSYNDEX, PsycArticles
EBSCOhost	"scale anchor"	All	1974-2021	ERIC, MEDLINE, PsycInfo, PSYNDEX, PsycArticles
EBSCOhost	"price anchor"	All	1974-2021	ERIC, MEDLINE, PsycInfo, PSYNDEX, PsycArticles
EBSCOhost	"anchor effect"	All	1974-2021	ERIC, MEDLINE, PsycInfo, PSYNDEX, PsycArticles
EBSCOhost	"anchor precision"	All	1974-2021	ERIC, MEDLINE, PsycInfo, PSYNDEX, PsycArticles
EBSCOhost	anchor moderator	All	1974-2021	ERIC, MEDLINE, PsycInfo, PSYNDEX, PsycArticles
OSF Preprints	Anchor	Economics, Psychology, Sociology, Psychiatry and Psychology	All	-
OSF Preprints	Anchoring	Economics, Psychology, Sociology, Psychiatry and Psychology	All	-
Web of Science	"anchoring" OR "anchoring effect"	Psychology, Psychology Applied, Psychology Experimental, Psychology Multidisciplinary, Psychology Social	1980-2021	SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH
Web of Science	"scale anchor"	All	1980-2021	SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH
Web of Science	"price anchor"	All	1980-2021	SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH
Web of Science	"anchor effect"	All	1980-2021	SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH
Web of Science	"anchor precision"	All	1980-2021	SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH
Web of Science	anchor moderator	All	1980-2021	SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH

Note. All searches were conducted between March 19 and March 31, 2021.

(i.e., the difference between the estimate and the anchor). Second, as recommended by Cheek and Norem (2018), we computed absolute adjustment from the anchor (i.e., the absolute difference between the estimate and the anchor). Third, we computed a score between 0 and 1 (0-1 score) depending on whether participants estimated the anchor or the true value (i.e., the difference between the estimate and the anchor divided by the difference between the true value and the anchor). Fourth, we computed a restricted 0-1 score, that is, trials with 0-1 scores below 0 or above 1 were coded 0 or 1, respectively (e.g., Yoon et al., 2021). Examples for hypothetical scenarios and adjustment- and 0-1-scores are presented in Table 2. Afterwards, the average anchoring effect size was computed as Hedges's g by comparing the average estimates from the high anchor group with those from the low anchor group. In cases in which anchors were continuous, correlations between anchors and estimates were computed and converted to Hedges's g.

For each study with multiple items and for each of the adjustment scores described above, we computed the average interitem correlations to estimate the reliability of the anchoring task. These correlations are related to commonly used coefficients such as Cronbach's α but independent of the number of items. Data

Figure 1

PRISMA Flow Chart for the Literature Search

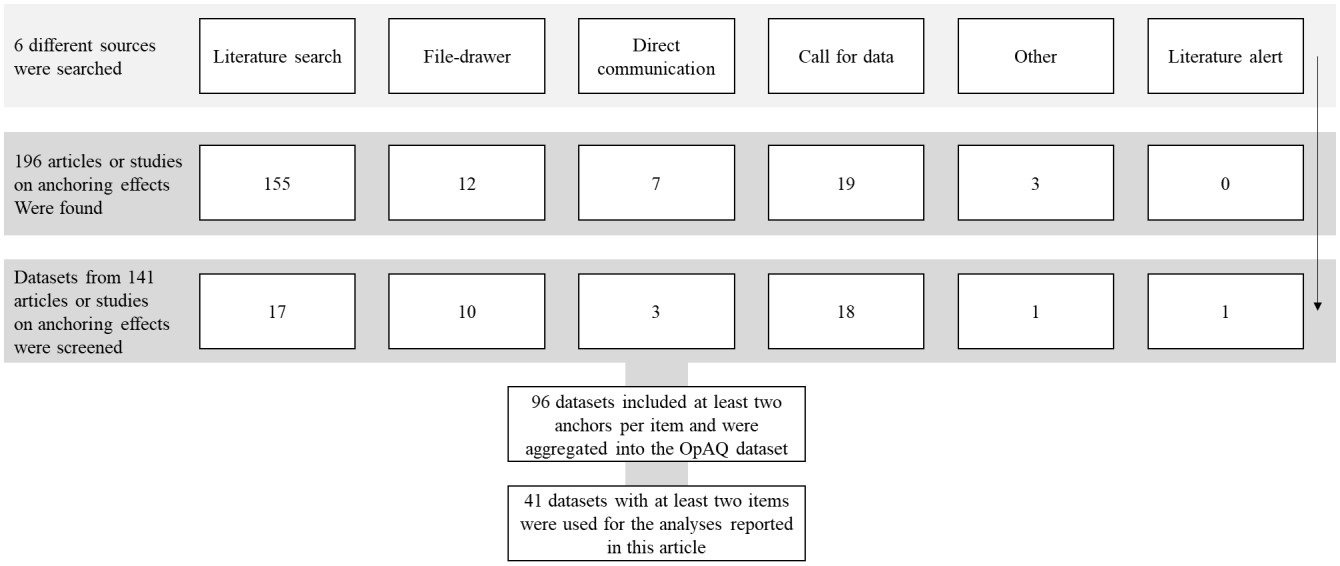


Table 2

Hypothetical anchoring scenarios with corresponding adjustment values and 0-1-Scores for two Persons A and B.

True Value	Anchor	Person A Estimate	Adjustment	0-1-Score	Person B Estimate	Adjustment	0-1-Score
5	6	5.1	-0.9	(5.1 – 6) / (5 – 6) = 0.9	5.6	-0.4	(5.6 – 6) / (5 – 6) = 0.4
5	4	4.9	0.9	(4.9 – 4) / (5 – 4) = 0.9	4.4	0.4	(4.4 – 4) / (5 – 4) = 0.4

Note. In this scenario, Person A is hardly influenced by the anchor as her estimate is much closer to the true value. Person B is more strongly influenced by the anchor.

were processed by one author and checked by a second one. Analyses were conducted using R version 4.1.1 (Team, 2018) and the packages apaTables (Stanley, 2021), ggplot2 (Wickham, 2016), gridExtra (Auguie, 2017), knitr (Xie, 2021), lmerTest (Kuznetsova et al., 2017), and xlsx (Dragulescu, 2014). All analyses were exploratory as most of the features of the data set could not be anticipated (e.g., sample size, number of data sets for which the interitem correlations could be computed, distribution of categories for moderator variables). The data and analysis scripts are available online (https://osf.io/ygnvb).

Results

The final data set that we used for the reliability analyses was derived from 41 studies, and a total sample size of N = 9,825 (Mdn sample size = 151, Mdn number of items = 6). For 32 of these studies, the 0-1 scores, which require true values for the anchoring questions, could be computed. The earliest study was from 2010 and the latest one from 2022. The overall mean interitem correlation of all 146 anchoring scores was an average interitem correlation of r = .126 (Mdn: r = .126; REML-estimate r = .137, 95% CI [.078, .197], p < .001). Of all 146 interitem correlations, 33 were

Table 3

Mean Interitem Correlations for the four Susceptibility Scores With Confidence Intervals

Susceptibility Score	Interitem correlation [95% CI]
Absolute Adjustment	0.107, [0.045, 0.17]
Adjustment	0.154, [0.092, 0.216]
Restricted 0-1 score	0.164, [0.100, 0.227]
0-1 score	0.131, [0.068, 0.195]

Note. Estimates and confidence intervals are based on a multilevel random-effects meta-analysis with susceptibility score nested in study.

larger than .3, which is considered desirable for reliable measurement (Hair et al., 2014, p. 123).[1] Distributions of interitem correlations for the four scores can be seen in Figure 2. The data set and analysis code are available online (https://osf.io/g95hp). There was no evidence of reporting biases (e.g., funnel plot asymmetry or differences between effect sizes from published and unpublished studies) in the dataset as discussed by Röseler and Schütz (2022). Interitem correlations from pre-prints did not differ from unpublished studies or studies that are part of a peer-reviewed journal article.

The highest interitem correlation was found by Lee and Morewedge (2022, Study 1a), where the adjustment score's mean interitem correlation was r = .933. Using a multilevel random-effects meta-analysis with susceptibility scores nested in studies, we compared average interitem correlations between the score types: Interitem correlations differed between scores, F(3, 103.43) = 4.12, p = 0.008, with interitem correlations for adjustment being the largest (mean r = .200) and the scores were correlated (see Figure 2 and Table 3 for descriptives and confidence intervals).

Discriminant Validity of Susceptibility Scores

We tested whether susceptibility scores were correlated with participants' mean estimates by study. For the sake of brevity, we provide results for the absolute adjustment and 0-1 scores only. Studies with high average interitem correlations had high correlations with estimates. That is, whether people estimated larger numbers was positively associated with the susceptibility scores. There were two exceptions with high interitem correlations and low correlations with mean estimates: First, the susceptibility scores in Study 1a by Lee and Morewedge (2022) were correlated and had relatively low correlations with mean estimates. This is probably because participants indicated their willingness to pay for three hotels at once for the same anchor. Second, the 0-1 score from Röseler et al. (2022) study had a low correlation with mean estimates, too, r = .167 (3 items) and a high interitem correlation, r = .749. Note that the interitem correlation of absolute adjustment was far lower (r = -.113). Among all studies, this difference between two scores was the largest. In this study, participants estimated the number of African members in the UN, the year the telephone was invented, and the maximum speed of a housecat. Scatterplots for the relationships are provided in Figure 3.

Impact of Study Features on Interitem Correlations

From a psychometric point of view, we considered anchoring effect size and number of different anchoring task types (e.g., height estimate, probability estimate) to be most important for the reliability. Larger effect sizes could be associated with higher interitem correlations, and heterogeneous tasks might be more eligible for mapping the construct of interest. However, correlational analyses and comparisons of means showed that the interitem correlations of the 0-1 scores were affected only by whether the direction of adjustment was known (despite adjustment direction being accounted for) and by whether a comparative question was included. Note that whenever there was a comparative question, we coded the adjustment direction as unknown. Both correlations were significant at α = .05 but should be interpreted with caution due to the large number of moderator tests and four dependent variables (i.e., the susceptibility scores). Scatterplots for the relationships between the interitem correlations, anchoring effect size, and task heterogeneity are presented in Figure 4. Among the multinomial variables, task type (i.e., speed, duration, proportion, distance, ...) and 0-1-score had the strongest association, F(3, 28) = 23.09, p = .012. Susceptibility scores for fixed anchors had the strongest interitem correlations, whereas those for subliminal/incidental anchors were very low (see Figure 5).

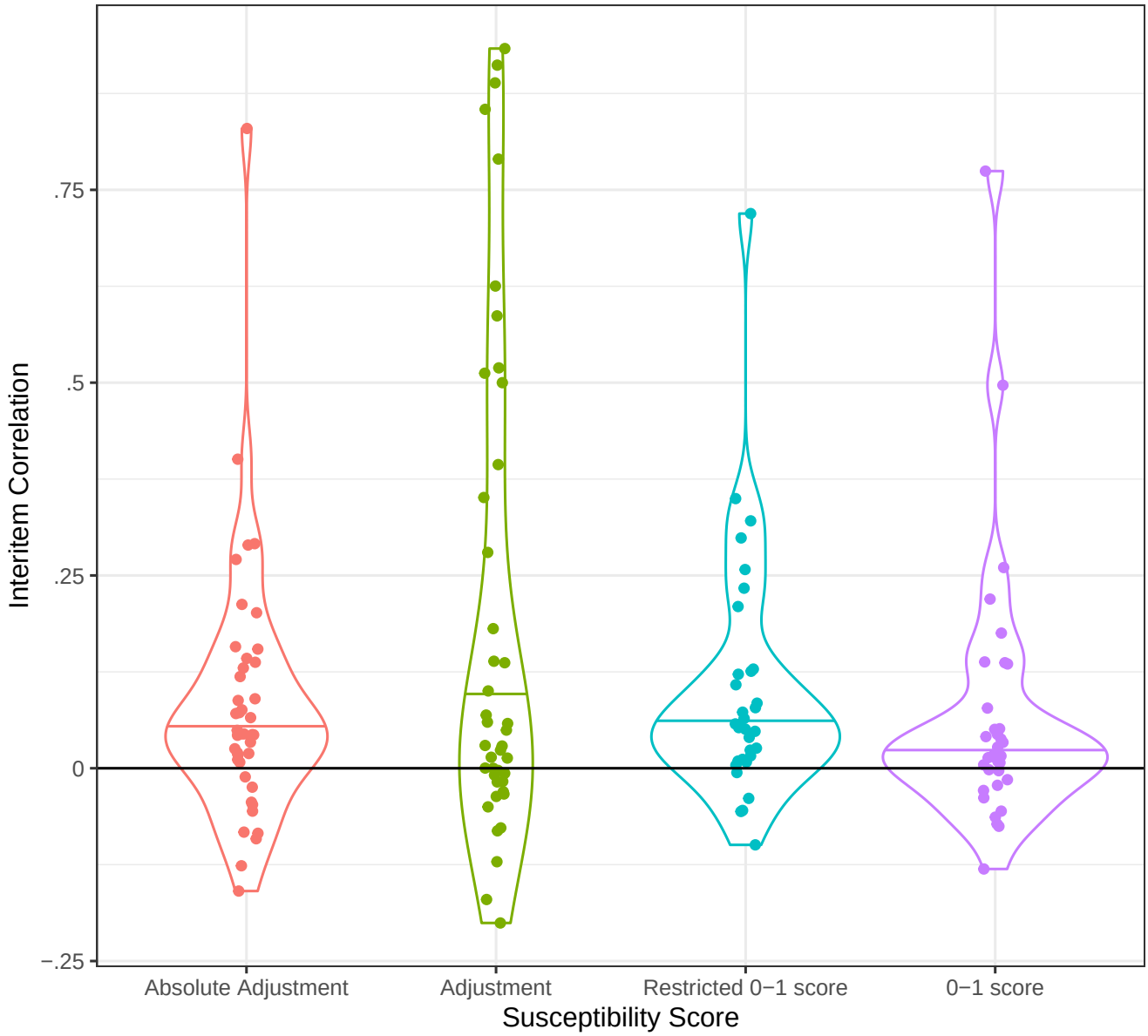
Discussion

Meta-analyses of interitem correlations from 41 studies showed that the mean interitem correlation of susceptibility to anchoring is very low and independent of the type of score, anchoring effect size, type of anchoring task, scale type, anchor type, and whether participants were incentivized or not. The few studies that had

[1] Note that, to achieve an acceptable Cronbach's Alpha (α > 0.7, George and Mallery, 2003, p. 231), r > 0.538 is necessary to use two items (which is the most frequent case apart from single-item measurement). r = .3 with two items corresponds to α = .462.

Figure 2

Violin Plots of Interitem Correlations by Susceptibility Score



Note. Each point represents one study's average interitem correlation for the respective susceptibility score. For each study, absolute adjustment and adjustment was computed. 0-1 scores and restricted 0-1 scores were computed only for studies with anchoring items that had a specific true value (e.g., height of the Eiffel tower) or a mean estimate from a group of participants who had not considered an anchor prior to providing their estimate.

high interitem correlations and thereby reliable measurements were plagued by serious concerns about validity. That is, if people seemed to show high susceptibility to anchoring, they in fact simply provided higher estimates on average. Such relationships result from paradigms that provide only high or only low anchors or do not vary anchors within participants across items. The presence of a comparative question (i.e., the question of whether the correct value is higher or lower than the anchor) had a large negative effect on the interitem correlations. Paradigm features such as the number of task types had no effect on the interitem correlations.

In sum, there is little evidence that susceptibility to anchoring can be measured reliably and validly. To our

Table 4

Means, Standard Deviations, and Correlations for the interitem correlations of anchoring items' raw estimates (1), interitem correlations of anchoring susceptibility scores (2-5), and study characteristics (6-13)

Variable	M	SD	1	2	3	4	5
1. Estimate	0.15	0.2					
2. Adjustment	0.21	0.32	0.69 [.48, .82]				
3. Absolute adjustment	0.08	0.17	0.74 [.56, .85]	0.46 [.18, .67]			
4. Score	0.08	0.21	0.52 [.20, .73]	0.43 [.09, .67]	0.56 [.26, .76]		
5. Restricted score	0.12	0.2	0.42 [.08, .67]	0.29 [-.07, .58]	0.47 [.15, .71]	0.93 [.86, .97]	
6. Anchoring effect size (Hedges's g)	0.7	0.52	0.01 [-.30, .32]	-0.01 [-.32, .31]	0.17 [-.15, .45]	0.11 [-.25, .45]	0.19 [-.18, .51]
7. Published?	0.43	0.5	-0.24 [-.51, .07]	-0.4 [-.63, -.10]	-0.16 [-.45, .15]	-0.19 [-.51, .17]	-0.19 [-.51, .17]
8. Proportion of women	0.61	0.15	-0.27 [-.57, .10]	-0.18 [-.51, .19]	-0.39 [-.66, -.03]	-0.2 [-.56, .24]	-0.21 [-.57, .22]
9. Mean age	30.23	7.9	0.04 [-.35, .42]	0.1 [-.30, .47]	0.15 [-.25, .51]	0.41 [-.00, .70]	0.5 [.11, .75]
10. Incentive	0.47	0.5	0.17 [-.15, .45]	-0.07 [-.37, .25]	0.17 [-.14, .45]	-0.04 [-.38, .31]	-0.06 [-.40, .30]
11. Direction	0.12	0.32	0.39 [.09, .63]	0.1 [-.22, .40]	0.59 [.34, .76]	0.23 [-.14, .54]	0.27 [-.09, .57]
12. Comparative question	0.66	0.48	-0.31 [-.58, .02]	-0.05 [-.37, .29]	-0.45 [-.68, -.14]	-0.09 [-.46, .30]	-0.21 [-.55, .18]

Note. Values in brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have produced the sample correlation (Cumming, 2014).

knowledge, we analyzed the most comprehensive set of data that could be found for assessing the reliability of people's susceptibility to anchoring.

Possible Reasons for the Reliability Problem

The lack of reliability in anchoring susceptibility scores can be due to reasons that are inherent in the construct or issues of measurement. Measurement-wise, susceptibility to anchoring may be obscured by numerous other traits that affect estimates but should be controlled for. For example, an extremely strong anchoring effect (i.e., a strong influence of the situation) could obfuscate the influence of a participant's own personality traits (cf. Cooper and Withey, 2009). Construct-wise, it is possible that anchoring affects everyone equally, and thus, there would not be individual differences in this aspect. Another possibility is that susceptibility is extremely volatile and changes rapidly. Finally, it is possible that there simply is no such construct.

The Problem With the Validity of Anchoring

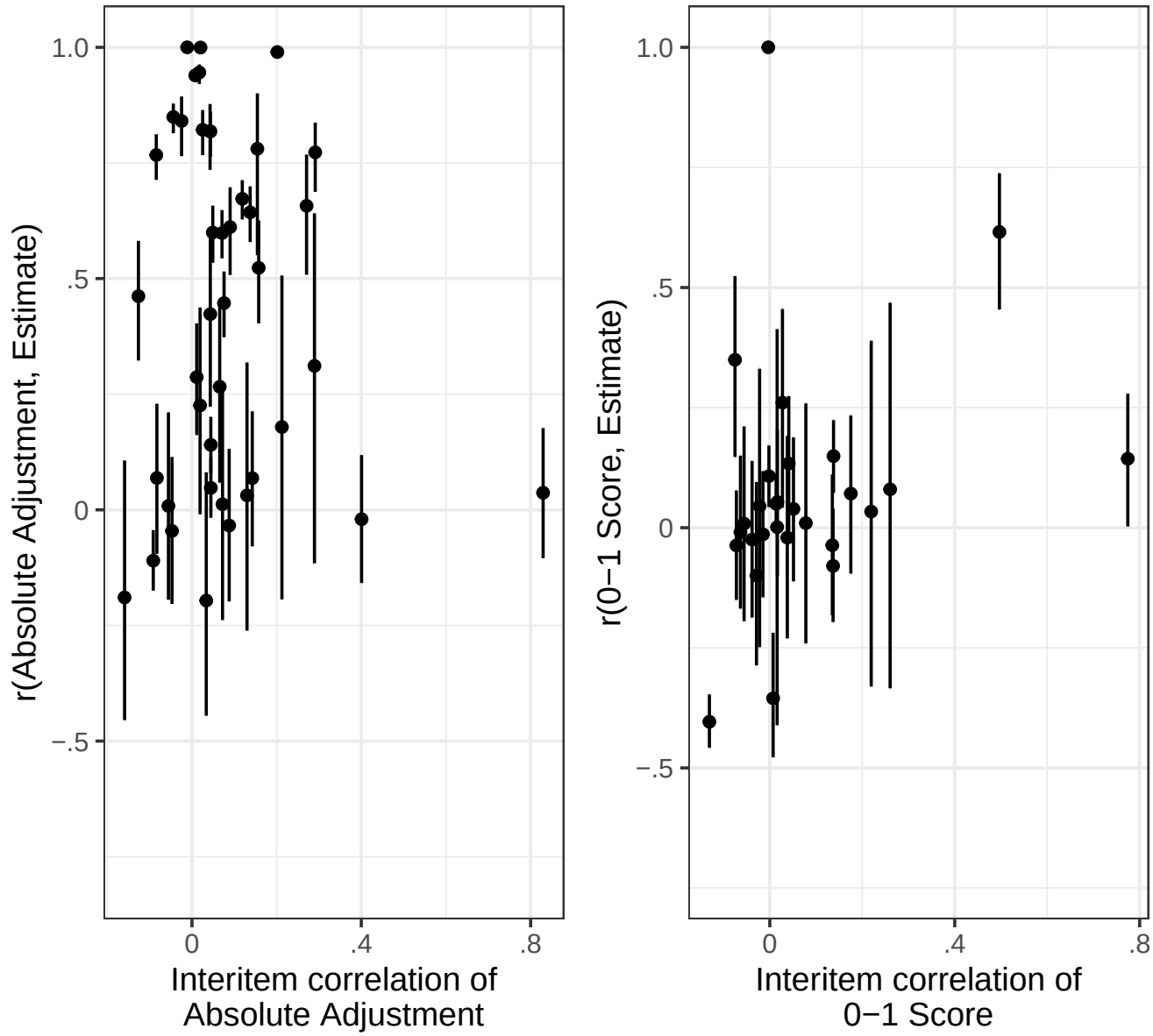
Like any measure, not only do measures of susceptibility to anchoring need to be reliable, but they also need to be valid. Interitem correlations should be large, and correlations with mean estimates should be as small as possible. In other words, how strongly participants adjust away from the anchor should be unrelated to whether they tend to provide larger or smaller estimates than others. For example, an anchoring paradigm might have two items with correct values of 10 and 50 and anchors of 20 and 100. There is likely to be a correlation between the absolute adjustment (i.e., the absolute difference between the anchor and the estimate) and an apparent susceptibility to anchoring in this case because both anchors are above the true values. In fact, according to our results, cases in which the reliability is acceptable suffer from high correlations between susceptibility scores and estimates. Thus, we conclude that these scores were not valid.

We noted that it is necessary to test high and low anchors for each participant to avoid responses that are confounded with response styles (i.e., whether people tend to estimate large or small numbers). However, this precaution is not sufficient because we still do not know at what level the anchors have to be. For example, if the true value is 50, should the anchors be 10 and 90 (true value $\pm$ 40) or 25 and 100 (true value $\times 2$

As the reliability of scores that claim to represent susceptibility to anchoring effects is very low, we recommend that claims of correlations between anchoring and personality should be taken with a grain of salt. There are contradictory findings for many moderators (e.g., Cheek and Norem, 2019 that resulted in null effects in meta-analyses, Röseler, 2021). Although desirable, a new and potentially reliable approach to measurement cannot be validated with classical paradigms because when susceptibility to anchoring is measured with classical paradigms, it apparently does not correlate reliably with anything. In other words, the black swan that we are looking for will essentially be black and not white like all the other swans. Thus, further attempts to determine associations between susceptibility to anchoring and personality traits require the development of new paradigms. Therefore, we advise researchers to report reliabilities, use as many and as heterogeneous items as possible, vary anchors within participants, use absolute adjustment or 0-1 scores, and think of new ways to measure susceptibility to anchoring beyond classical

Figure 3

Interitem correlations and Discriminant Validities for Absolute Adjustment and 0-1 Scores



anchoring paradigms.

Report Reliability

If a tendency is not measured reliably, it is unreasonable to expect that it will produce stable correlations with another tendency or trait. Still, there are numerous findings that indicate that personality moderates the susceptibility to anchoring, but they resulted from the use of paradigms that have been shown to provide low reliabilities. We recommend that these correlations be replicated in preregistered studies and ideally with more reliable paradigms. In this approach, researchers should also correlate susceptibility scores with response tendencies to alleviate concerns about validity. Most importantly, we recommend that researchers report the reliabilities of susceptibility to anchoring along with the reliabilities of potential personality moderators.

Figure 4

Scatterplots for Interitem Correlations by Anchoring Effect Sizes and Numbers of Different Task Types

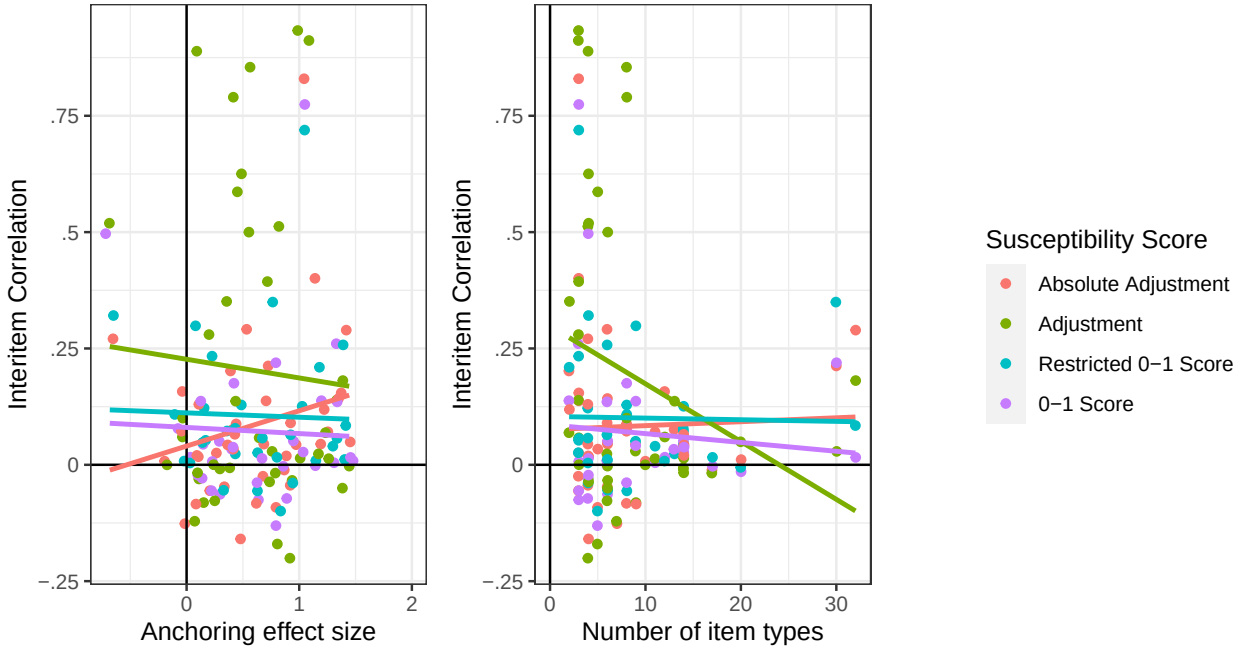
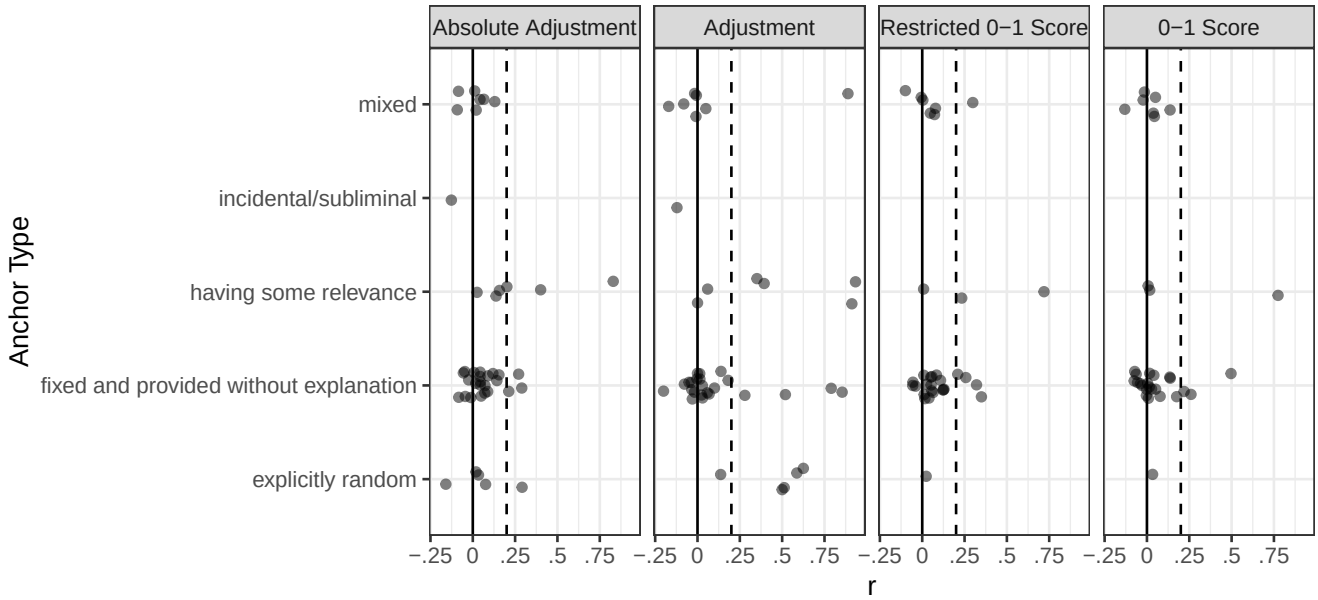


Figure 5

Scatterplots for Interitem Correlations by Type of Anchor



Use as Many and as Heterogeneous Items as Possible

Susceptibility to anchoring should be independent of item types; that is, if people are highly susceptible to anchoring, they should stick relatively close to the anchor when estimating whether the Rio Grande is longer or shorter than 5,000 miles but also when estimating whether the average number of babies born each day in the United states is more or fewer than 100. Past paradigms did not achieve such reliable measurement when items from different domains were used. Notably, if reliability was higher, then validity was lower because tasks were relatively homogeneous. Thus, in these cases, reliability may reflect task-specific expertise instead of susceptibility to anchoring.

Vary Anchors Within Participants

To control for the mean estimate, anchors that are above and below the potential unanchored estimates need to be implemented in the paradigm. Note that for our analyses, paradigms that did not have multiple anchors were excluded, but anchors could still be high only in cases where the true value was unknown (e.g., prices of products). If no anchoring effect has occurred, differences in adjustments from anchors might not reflect susceptibility to anchoring. Thus, for validation, we also recommend making sure that anchoring effects are actually present. If high and low anchors are used, this test is easy to establish.

Beware of Adjustment Scores

In our meta-analysis, we examined four different scores that are used to measure susceptibility. Although all of the scores were correlated, we advise against the use of adjustment scores (i.e., the difference between the estimate and the anchor) and restricted 0-1 scores (i.e., the difference between the estimate and the anchor divided by the difference between the true value and the anchor) because adjustment scores might not be valid if the direction of adjustment is unknown (Cheek & Norem, 2018), and restricted 0-1 scores discard information about the anchoring effect size. For example, scores above 1 indicate that no assimilation toward the anchor occurred. Instead, we recommend absolute adjustment and 0-1 scores, which are moderately intercorrelated and not subject to these problems. Note that although 0-1 scores make comparisons between anchoring items easier, true values must be used in the questions.

Our last recommendation, which also ties into a limitation of our work, is to use explicit models of how susceptibility to anchoring is related to participants' estimates. The scores that we used are based on simple models of anchoring, in which estimates can be regarded as the weighted mean of the true value and the anchor plus an error. However, non-linear models that account for the nonlinear influence of extreme anchors (e.g., Chapman and Johnson, 1994; Fechner, 1860) are likely more precise.

Limitations

A large proportion of the data upon which our conclusions were based came from openly accessible data sets that were published after 2015. These data might not be representative of anchoring research with respect to moderator variables. Despite numerous calls for data and correspondence with the authors of data sets that were not openly accessible, only a few researchers provided us with their data sets. Given that open data requirements were only recently implemented in the field, we believe that data from future research will be easier to add to the data set.

By coding more than 10 potential moderator variables that were included in other meta-analyses of anchoring effects (e.g., Bystranowski et al., 2021; Li et al., 2021), we aimed to obtain a comprehensive overview of the effects of different paradigms and study features. Certainly, moderators that are still unknown could account for the large degree of heterogeneity in anchoring effects. Such research may be an avenue for future endeavors. Luckily, additional moderators (e.g., the precision of the anchor) can be computed from our data set, which is openly available at https://osf.io/ygnvb.

Our analyses of reliabilities were not preregistered as is the case for most current meta-analyses. Sample size planning (i.e., predicting how many openly accessible data sets we would find and the extent to which other researchers would be willing to share their data) was very difficult. Also, we did not know a priori which tests were feasible. For example, although we coded whether participants were lay people or experts, there was no data set with expert estimates, which is why this moderator could not be analyzed. We want to emphasize that all analyses were exploratory, and the addition of further data will provide information about the robustness of our results.

Conclusion

We consider mapping and solving the reliability problem to be vital for the progress of the field of anchoring research. The integration of or discrimination between different phenomena can hardly proceed without solving the reliability problem first. With our meta-analysis of open anchoring data sets, we provide evidence that reliable measurement of the susceptibility to anchoring

has been achieved only in rare cases—and that these cases lack validity.

There are three directions where researchers can go from this point: The first is to separate anchoring research from personality research. Then, working with approaches such as the scale distortion theory (Frederick & Mochon, 2012) could be promising. However, in following this path, integrating different but theoretically linked phenomena, such as anchoring and hindsight bias (see also Pohl et al., 2003), would not be easy due to the lack of reliability. The second is to dissolve the boundaries of anchoring phenomena by creating new paradigms that yield effects that follow the same line of interpretation on the basis of different variables (e.g., Bahník, 2021a; Frederick and Mochon, 2012). For example, some researchers have been trying to integrate mouse tracking into the anchoring paradigm. Another option is to use choices between numbers instead of numeric estimates to allow for more fine-grained modeling approaches, such as drift-diffusion models (e.g., Hedge et al., 2018, p. 1181). Third, researchers may create formal models of anchoring (e.g., Pohl et al., 2003; Turner and Schley, 2016) that can serve as the basis for claims of reliable measurement. Such new approaches may have the potential to shed light on the generalizability of anchoring phenomena but also on the validity of existing theoretical accounts.

Acknowledgments

We thank the University of Bamberg's CatchUp+ program, aimed at supporting researchers with children during the Covid-19 pandemic, for funding this project. We thank Jane Zagorski for language editing.

CRediT Author Statement

- LR: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing
- LW: Data curation, Investigation, Resources, Software, Visualization, Writing – original draft, Writing – review & editing
- KH: Validation, Writing – review & editing
- ES: Data curation, Software, Writing – review & editing
- MG: Data curation, Writing – review & editing
- PT: Data curation, Writing – review & editing
- FSW: Investigation, Writing – review & editing
- AS: Funding acquisition, Resources, Supervision, Writing – review & editing

Author Contact

Correspondence concerning this article should be addressed to Lukas Röseler, lukas.roeseler@uni-muenster.de.

ORCID IDs: LR: 0000-0002-6446-1901 , KACH: 0009-0003-9679-8263 , PT: 0000-0002-4701-4309 , AS: 0000-0002-6358-167X

Conflict of Interest and Funding

The authors declare that they have no conflict of interest. This research was partly supported by funding from the University of Bamberg's CatchUp+ program awarded to Lukas Röseler. The funding source was not involved in the study design, data collection, data analysis, interpretation of data, writing, or decision to submit.

Open Science Practices

This article earned the Open Data, Open Materials, and Open Code badge for making the data, materials, and code openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Auguie, B. (2017). Gridextra: Miscellaneous functions for "grid" graphics [R package version 2.3]. https : / / CRAN . R - project . org / package = gridExtra

Bahník, Š. (2021a). Anchoring does not activate examples associated with the anchor value. *Advance online publication*. https://doi.org/10.31234/osf.io/4j5wb

Bahník, Š. (2021b). Anchoring without scale distortion. *Judgment and Decision Making*, *16*(1), 131. https://doi.org/10.31234/osf.io/2q8hj

Bergman, O., Ellingsen, T., Johannesson, M., & Svensson, C. (2010). Anchoring and cognitive ability. *Economics Letters*, *107*(1), 66–68. https://doi.org/10.1016/j.econlet.2009.12.028

Berthet, V. (2021). The measurement of individual differences in cognitive biases: A review and improvement. *Frontiers in Psychology*, *12*, 630177. https://doi.org/10.3389/fpsyg.2021.630177

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bystranowski, P., Janik, B., Próchnicki, M., & Skórska, P. (2021). Anchoring effect in legal decision-making: A meta-analysis. *Law and Human Behavior*, *45*(1), 1–23. https://doi.org/10.1037/lhb0000438

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making*, *7*(4), 223–242. https://doi.org/10.1002/bdm.3960070402

Cheek, N. N., & Norem, J. K. (2018). On moderator detection in anchoring research: Implications of ignoring estimate direction. *Collabra: Psychology*, *4*(1), 12. https://doi.org/10.1525/collabra.125

Cheek, N. N., & Norem, J. K. (2019). Are big five traits and facets associated with anchoring susceptibility? *Social Psychological and Personality Science*, *9*(2), 194855061983700. https://doi.org/10.1177/1948550619837001

Cheek, N. N., & Norem, J. K. (2022). Individual differences in anchoring susceptibility: Verbal reasoning, autistic tendencies, and narcissism. *Personality and Individual Differences*, *184*, 111212. https://doi.org/10.1016/j.paid.2021.111212

Cooper, W. H., & Withey, M. J. (2009). The strong situation hypothesis. *Personality and Social Psychology Review*, *13*(1), 62–72. https://doi.org/10.1177/1088868308329378

Dragulescu, A. A. (2014). Xlsx: Read, write, format excel 2007 and excel 97/2000/xp/2003 files [R package version 0.5.7]. https://CRAN.R-project.org/package=xlsx

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396. https://doi.org/10.1111/1467-9280.00372

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Fechner, G. T. (1860). *Elemente der psychophysik*. Breitkopf und Härtel.

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, *141*(1), 124–133. https://doi.org/10.1037/a0024006

George, D., & Mallery, P. (2003). *Spss for windows step by step: A simple guide and reference ; 11.0 update* (4th). Allyn; Bacon.

Gertner, A., Zaromb, F., Schneider, R., & Metthews, G. (2016). *The assessment of biases in cognition: Development and evaluation of an assessment instrument for the measurement of cognitive bias* (tech. rep. No. MTR160163). The MITRE Corporation. McLean, VA.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th). Pearson Education Limited.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science*, *19*(2), 121–127. https://doi.org/10.1111/j.1467-9280.2008.02057.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lee, C.-Y., & Morewedge, C. K. (2022). Noise increases anchoring effects. *Psychological Science*, *33*(1), 60–75.

Li, L., Maniadis, Z., & Sedikides, C. (2021). Anchoring in economics: A meta-analysis of studies on willingness-to-pay and willingness-to-accept. *Journal of Behavioral and Experimental Economics*, *90*, 101629. https://doi.org/10.1016/j.socec.2020.101629

Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, *122*(1), 69–79. https://doi.org/10.1016/j.obhdp.2013.04.002

Mussweiler, T., & Strack, F. (1999). Comparing is believing: A selective accessibility model of judgemental anchoring. *European Review of Social Psychology*, *10*(1), 135–167. https://doi.org/10.1080/14792779943000044

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Consid-

ering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142–1150. https://doi.org/10.1177/01461672002611010

Parsons, S., Kruijt, A.-W., & Fox, E. (2018). Psychological science needs a standard practice of reporting the reliability of cognitive behavioural measurements [Advance online publication]. https://doi.org/10.17605/OSF.IO/6KA9Z

Pohl, R. F., Eisenhauer, M., & Hardt, O. (2003). Sara: A cognitive process model to simulate the anchoring effect and hindsight bias. *Memory*, *11*(4–5), 337–356. https://doi.org/10.1080/09658210244000487

Röseler, L. (2021). *Anchoring effects: Resolving the contradictions of personality moderator research* [Doctoral dissertation, University of Bamberg]. https://doi.org/10.20378/irb-49951

Röseler, L., Bögler, H. L., Koßmann, L., Krueger, S., Bickenbach, S., Bühler, R., della Guardia, J., Köppel, L.-M. A., Ponader, S., Roßmaier, K., et al. (2022). Replicating epley and gilovich: Need for cognition, cognitive load, and forewarning do not moderate anchoring effects. *Advance online publication*.

Röseler, L., & Schütz, A. (2022). Hanging the anchor off a new ship: A meta-analysis of anchoring effects. *Advance online publication*. https://doi.org/10.31234/osf.io/wf2tn

Röseler, L., Schütz, A., & Starker, U. (2019). Cognitive ability does not and cannot correlate with susceptibility to anchoring effects. *Advance online publication*. https://doi.org/10.31234/osf.io/bnsx2

Schindler, S., Querengässer, J., Bruchmann, M., Bögemann, N. J., Moeck, R., & Straube, T. (2021). Bayes factors show evidence against systematic relationships between the anchoring effect and the big five personality traits. *Scientific Reports*, *11*(1), 7021. https://doi.org/10.1038/s41598-021-86429-2

Schultze, T., Gerlach, T. M., & Rittich, J. C. (2018). Some people heed advice less than others: Agency (but not communion) predicts advice taking. *Journal of Behavioral Decision Making*, *31*(3), 430–445. https://doi.org/10.1002/bdm.2065

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2017). On the inability to ignore useless advice. *Experimental Psychology*, *64*(3), 170–183. https://doi.org/10.1027/1618-3169/a000361

Smeesters, D., Yzerbyt, V. Y., Corneille, O., & Warlop, L. (2009). When do primes prime? the moderating role of the self-concept in individuals' susceptibility to priming effects on social behavior. *Journal of Experimental Social Psychology*, *45*(1), 211–216. https://doi.org/10.1016/j.jesp.2008.09.002

Stanley, D. (2021). Apatables [Version R package version 2.0.8]. https://CRAN.R-project.org/package=apaTables

Team, R. C. (2018). R [R Foundation for Statistical Computing. Vienna, Austria]. https://www.R-project.org/

Teovanović, P. (2019). Individual differences in anchoring effect: Evidence for the role of insufficient adjustment. *Europe's Journal of Psychology*, *15*(1), 8–24. https://doi.org/10.5964/ejop.v15i1.1691

Thorsteinson, T. J. (2011). Initiating salary discussions with an extreme request: Anchoring effects on initial salary offers. *Journal of Applied Social Psychology*, *41*(7), 1774–1792. https://doi.org/10.1111/j.1559-1816.2011.00779.x

Townson, C. D. (2019). *The anchoring effect: A meta-analysis* [Doctoral dissertation, Michigan State University] [ProQuest Dissertations Theses Global].

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665. https://doi.org/10.1177/1745691614552498

Turner, B. M., & Schley, D. R. (2016). The anchor integration model: A descriptive model of anchoring effects. *Cognitive Psychology*, *90*, 1–47. https://doi.org/10.1016/j.cogpsych.2016.07.003

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (2nd). Springer.

Xie, Y. (2021). Knitr [Version R package version 1.36]. https://cran.r-project.org/web/packages/knitr/

Yoon, H., Scopelliti, I., & Morewedge, C. K. (2021). Decision making can be improved through observational learning. *Organizational Behavior and Human Decision Processes*, *162*, 155–188. https://doi.org/10.1016/j.obhdp.2020.10.011